

(19) KOREAN INTELLECTUAL PROPERTY OFFICE

KOREAN PATENT ABSTRACTS

(11)Publication number: 20010106666 A
 (43)Date of publication of application: 07.12.2001

(21)Application number: 20000027518

(22)Date of filing: 22.05.2000

(71)Applicant: GEONJISOFT CO., LTD.
TRAVELHOW.COM, INC.(72)Inventor: KANG, SEONG GU
YOO, HONG JIN

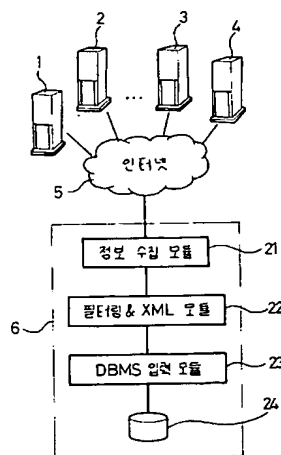
(51)Int. Cl. G06F 17/40

(54) METHOD AND SYSTEM FOR EXTRACTING INFORMATION FROM WEB PAGE AND STORING EXTRACTED INFORMATION

(57) Abstract:

PURPOSE: A method and system for extracting and storing information are provided to extract information from a web page in a processing a portion having a specific pattern out of illogical structures of an HTML document, and to store the extracted information.

CONSTITUTION: A server(6) is connected with a plurality of client PCs(1,2) and a plurality of servers (3,4) and provided for supplying a specific service. The server(6) has a program for being operated in accordance with a specific object, and performs a work requested by the client PCs(1,2) rapidly. The client PCs(1,2) are operated as a method of receiving data displaying a web page and performing a necessary work and returning data to the sever(6). The server(6) may include an information collecting module(21) for collecting information from a web page including a specific subject out of many subjects on the Internet, a filtering and XML module(22) for extracting information from information collected by the information collecting module(21) as an available form, and a DBMS input module(23) for storing data outputted from the filtering and XML module(22) in a recording medium(24).



&copy; KIPO 2002

Legal Status

Date of request for an examination (20000522)

Notification date of refusal decision (20021021)

Final disposal of an application (rejection)

Date of final disposal of an application (20021021)

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(51) Int. Cl.

G06F 17/40

(11) 공개번호 특2001-0106666

(43) 공개일자 2001년12월07일

(21) 출원번호 10-2000-0027518

(22) 출원일자 2000년05월22일

(71) 출원인 주식회사 건지소프트 복인근

전북 전주시 덕진구 덕진동1가 664-14주식회사 트레이블하우닷컴 강성구

서울 종로구 관훈동 198-8 종로빌딩 9층

(72) 발명자 유홍진

경기도시흥시정왕동금강아파트117동803호

강성구

경기도남양주시도농동127-4

(74) 대리인 이정익

심사청구 : 있음

(54) 웹페이지로부터 정보를 추출하고 저장하기 위한 방법과시스템, 그리고 추출된 데이터를 저장하는 저장매체

요약

본 발명은 웹상의 HTML문서로부터 정보를 수집하는 방법 및 시스템에 관한 것으로서, 특히 웹상에 존재하는 HTML문서에 대한 정보를 수집하고, HTML문서로부터 패턴에 따라 필요한 데이터를 추출하며, 데이터베이스에 저장하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법 및 시스템, 저장매체에 관한 것이다.

본 발명은, 웹상의 HTML 문서의 링크정보를 수집하여 문서의 정보를 구축하는 제1단계와, 수집된 HTML 문서의 링크정보를 이용하여 특정태그를 기준으로 HTML문서를 분석하는 제2단계와, 분석된 HTML문서에 대한 HTML문서의 패턴을 정의하는 제3단계와, 상기 제3단계에서 정의된 패턴에 따라서 데이터를 추출하는 제4단계와, 추출된 데이터를 데이터베이스로 저장하는 제5단계를 포함하는 방법이며, 이것을 구현하기 위한 시스템, 그리고 저장매체를 포함한다.

본 발명에 의하면 HTML문서의 비논리적인 구조중에서 특정한 패턴(pattern)을 가진 부분을 처리함에 있어서 범용성을 가질 수 있으며, 패턴분석을 통하여 추출되는 정보를 논리적인 구조로 추출해냄으로써 보다 유용하게 정보를 처리할 수 있으며, 데이터베이스와의 연동을 통하여 정보의 재활용을 용이하게 할 수 있다.

도표도

도1

색인어

웹페이지, HTML가공, XML, 정보, 추출, 패턴분석, 태그, DB저장, 저장매체

명세서

도면의 간단한 설명

도 1은 본 발명의 서버시스템이 포함된 인터넷 연결상태도.

도 2는 본 발명의 방법을 표시하는 흐름도.

도 3a 및 도 3b는 본 발명에 의한 HTML 문서 구조분석방식을 나타내는 도면.

도 4는 본 발명에 의한 데이터추출방식을 나타내는 도면.

도 5는 HTML문서로부터 추출된 정보를 활용한 경매와 광고시스템과의 연계방식을 설명하는 일실시예의 구성도.

도 6은 본 발명에 의한 사용자인터페이스의 일실시예.

도 7a 및 도 7b는 본 발명을 이용한 HTML문서의 분석예.

<도면의 주요 부분에 대한 부호의 설명>

- | | |
|------------------|-----------------|
| 1: 클라이언트PC | 2: 클라이언트PC |
| 3: 서버 | 4: 서버 |
| 5: 인터넷 | 6: (정보수집용)서버 |
| 21: 정보수집모듈 | 22: 필터링 & XML모듈 |
| 23: DBMS입력모듈 | 24: 저장매체 (상품DB) |
| 25: 사용자DB | 26: 검색인터페이스 |
| 27: 사용자관리수단 | 28: 검색수단 |
| 29: 경매관리수단 | 30: 광고 및 푸시관리수단 |
| 31: 푸시, 이메일서비스수단 | |

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 웹상의 HTML문서로부터 정보를 수집하는 방법 및 시스템에 관한 것으로서, 특히 웹상에 존재하는 HTML문서에 대한 정보를 수집하고, HTML문서로부터 패턴에 따라 필요한 데이터를 추출하여, 데이터베이스에 저장하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법 및 시스템, 저장매체에 관한 것이다.

인터넷은 종종 무한한 정보의 바다라는 인식과 함께 불필요한 정보가 넘치는 쓰레기의 바다로 비유된다. 이러한 인식의 근본적인 원인은 기하급수적으로 증가하는 웹페이지를 일일이 분석하고 필요한 정보를 추출하는데 많은 시간과 비용이 소요되기 때문에, 정작 고부가가치를 가진 정보를 획득하는데 많은 시간과 비용이 들기 때문이다.

현재 인터넷에서 사용되고 있는 표준문서포맷인 HTML은 사용하기 용이한 장점때문에 인터넷의 커다란 발전을 초래하였으나, 이로 인하여 일반인들이 접근할 수 있는 정보의 폭발적인 증가를 야기함으로써 결과적으로 사용자들이 적시에 필요한 정보를 이용하기 어렵다는 문제점을 발생시키고 있다.

이러한 문제의 주요인은 HTML문서가 논리적인 구조를 표현하기 보다 문서의 외형을 표현하는데 중점을 두고 개발되었기 때문에 기인하는 것이다. 따라서 웹이용자들은 문서의 생성보다 생성된 문서로부터 유용한 정보를 효율적으로 추출할 수 있는 방법의 추구에 주력하고 있다.

이러한 문제점을 해결하기 위하여 웹페이지로부터 필요한 정보를 추출하기 위한 풀텍스트기반의 검색엔진이 등장하여 사용되고 있다. 풀텍스트(fulltext)란 웹페이지의 본문전체에 걸쳐서 용어를 검색하여 추출하는 방식으로, 어느 정도 정보검색에 대한 욕구를 충족시켜주고 있으나, 지나치게 많은 검색결과가 출력되며 검색에 많은 시간이 소요되는 문제점이 발생되고 있다. 또한 출력된 검색결과로부터 수작업을 통하여 유용한 정보를 취사, 선택하여야 하는 등의 한계점이 노출되고 있다.

따라서 HTML문서의 논리적 구조의 한계점을 보다 논리적인 구조로 변환할 수 있는 XML의 새로운 문서 표준안이 제창되었으나, 현재까지 웹페이지의 대부분을 차지하는 HTML문서에 대한 정보수집이 과제로 남아 있는 것이다.

발명이 이루고자하는 기술적 과제

본 발명은 상기와 같은 문제점을 해결하기 위하여 안출된 것으로서, 본 발명의 목적은 HTML문서의 비논리적인 구조 중에서 특정한 패턴(pattern)을 가진 부분을 처리함에 있어서 범용성을 가지는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법 및 시스템을 제공하는데 있다.

본 발명의 다른 목적은, 패턴분석을 통하여 추출되는 정보를 논리적인 구조로 추출해냄으로써 보다 유용하게 정보를 처리할 수 있는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법 및 시스템을 제공하는데 있다.

본 발명의 다른 목적은, 데이터베이스와의 연동을 통하여 정보의 재활용을 용이하게 할 수 있는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법 및 시스템을 제공하는데 있다.

본 발명의 또다른 목적은 웹페이지로부터 정보를 추출하고 저장하기 위한 본 발명의 방법을 이용하여 다양한 형태의 사업성있는 서비스모형을 제공함으로써 정보획득의 신속, 정확성을 제고하여 경쟁력을 높이는 데 있다.

발명의 구성 및 작용

상기 목적을 달성하기 위한 본 발명의 웹페이지로부터 정보를 추출하고 저장하기 위한 방법은, 웹상의 HTML 문서의 링크정보를 수집하여 문서의 정보를 구축하는 제1단계와, 수집된 HTML문서의 링크정보를 이용하여 특정태그를 기준으로 HTML문서를 분석하는 제2단계와, 분석된 HTML문서에 대한 HTML문서의 패턴을 정의하는 제3단계와, 상기 제3단계에서 정의된 패턴에 따라서 데이터를 추출하는 제4단계와, 추출된 데이터를 데이터베이스형태로 저장하는 제5단계를 포함한다.

본 발명의 다른 특징에 의하면, 패턴필터링에 의하여 추출된 데이터를 XML(eXtensible Markup Language)

문서로 변환하는 단계를 더 포함한다.

또한 본 발명의 시스템은, 인터넷에 산재한 웹페이지중의 특정주제를 포함한 웹페이지로부터 정보를 수집하는 정보수집모듈(21)과, 상기 정보수집모듈(21)에 의하여 수집된 정보로부터 유용한 형태로 정보를 추출하기 위한 필터링 & XML 모듈(22)과, 상기 필터링 & XML 모듈(22)로부터 출력되는 데이터를 특정포맷으로 변환하여 저장매체(24)에 저장하기 위한 DBMS 입력모듈(23)을 포함한다.

이하 첨부된 도면을 참고하여 본 발명을 상세히 설명하면 다음과 같다.

도 1은 종래 인터넷 연결상태를 도시한 개략도로서, 다수의 클라이언트PC(1,2,...)들과, 다수의 서버(3,4,...)들이 연결되어 있으며, 본 발명의 방법에 의한 특정 서비스를 제공하기 위한 고성능컴퓨터인 서버(6)가 연결된다. 서버(6)는 특정목적에 맞게 작동되도록 프로그램되어 있어서 신속하게 클라이언트PC가 요청하는 작업을 수행한다. 본 발명의 방법도 역시 서버(6)에 의하여 수행되며, 서버는 적어도 듀얼 펜티엄 III 이상의 중앙처리장치를 구비한 컴퓨터와, 대용량저장매체를 이용하는 것이 바람직하다.

서버와 클라이언트PC의 차이는 능동적으로 웹페이지를 제공하는가 아닌가하는 점에서 차이가 있다. 클라이언트PC는 서버로부터 웹페이지를 표시하는 데이터를 전송받아서 필요한 작업을 하고, 다시 서버로 데이터를 돌려보내는 방식으로 작동된다. 설명의 편의를 위하여 2대의 클라이언트PC와 2대의 서버만을 표시하였지만, 웹상에는 수 백만대의 서버와 수 천만대의 클라이언트PC가 연결되어 있으며, 그 숫자는 기하급수적으로 증가하고 있음이 통계에 의하여 알려져 있다.

본 발명의 방법이 수행되는 일실시예로서, 서버(6)의 구성은, 인터넷에 산재한 웹페이지중의 특정주제를 포함한 웹페이지로부터 정보를 수집하는 정보수집모듈(21)과, 상기 정보수집모듈(21)에 의하여 수집된 정보로부터 유용한 형태로 정보를 추출하기 위한 필터링 & XML 모듈(22)과, 상기 필터링 & XML 모듈(22)로부터 출력되는 데이터를 저장매체(24)에 저장하기 위한 DBMS 입력모듈(23)을 포함할 수 있다.

상기 정보수집모듈(21)은 인터넷에 산재한 서버(3,4,...)들에 저장되어 있는 웹페이지에 대한 데이터를 전송받아서 저장하는 역할을 한다. 정보수집모듈(21)은 웹로봇의 한 형태로서 사용자가 입력한 시작 URL을 기점으로 하여 웹사이트의 링크(link)를 따라가면서 HTML문서의 링크정보를 수집한다.

상기 필터링 & XML 모듈(22)은 상기 정보수집모듈(21)에 의하여 수집된 웹페이지데이터를 분석하고 필요한 데이터를 추출하기 위한 것으로서, 분석단계에서는 태그(tag)를 이용하며, 추출단계에서는 문서패턴의 정의규칙을 통하여 데이터를 필요한 형태로 가공하는 역할을 수행한다. 정의규칙은 데이터를 필요한 형태로 가공하기 위한 변환규칙을 의미하며, 후에 상술된다. 상기 분석단계와 추출단계에 의하여 필터링된 데이터는 또한 필요한 경우에 XML타입으로 데이터를 변환할 수 있다.

상기 DBMS 입력모듈(23)은 상기 분석단계 및 추출단계를 거치면서 필터링된 데이터를 저장매체(24)에 저장하는 기능을 수행한다. 저장매체(24)는 예를 들어 하드디스크, 플로피디스크, 광자기디스크, 플래시메모리 등과 같이 데이터의 저장, 삭제가 가능하며, 특히 저장속도가 빠른 매체를 사용하는 것이 바람직하다. 저장매체(24)에 저장된 데이터들은 별도의 기능모듈들을 더 추가함으로써 데이터제공서비스의 이용될 수 있으며, 예를 들면 여행정보검색등에 사용된다.

도 2는 상술한 시스템에서 수행되는 본 발명의 방법을 도시한 흐름도이다.

본 발명의 방법은, 웹상의 HTML 문서의 링크정보를 수집하여 문서의 정보를 구축하는 제1단계(S1)와, 수집된 HTML문서의 링크정보를 이용하여 특정태그를 기준으로 HTML문서를 분석하는 제2단계(S2)와, 분석된 HTML문서에 대한 HTML문서의 패턴을 정의(선언)하는 제3단계(S3)와, 상기 제3단계(S3)에서 정의된 패턴에 따라서 데이터를 추출하는 제4단계(S4)와, 추출된 데이터를 데이터베이스형태로 저장하는 제5단계(S5)를 포함하고 있다. 또한 필요에 따라서 추출된 데이터를 XML로 변환하는 단계(S4a)를 더 포함할 수 있다.

상기 제1단계(S1)는 HTML문서 필터링 기법을 이용하여 웹상의 HTML문서의 정보를 구축하는 단계로서, 웹로봇(도 1의 정보수집모듈(21))은 사용자가 입력한 시작 URL을 기점으로 하여, 웹사이트의 링크(웹페이지에 표시될)를 따라가며 HTML 문서의 링크정보를 수집한다. 수집된 정보는 서버(6)에 저장된다. 시작 URL은 사용자가 관심있는 분야에 대한 정보를 획득하기 위하여 해당 분야에 대한 내용을 포함하는 임의의 URL이 될 수 있다. 예를 들어서 여행에 관한 정보를 수집하고 싶으면, 여행정보제공사이트 또는 여행사의 웹사이트 주소를 입력시킨다.

상기와 같이 웹상의 HTML문서의 정보가 구축된 후에, 제2단계(S2)로서, HTML문서의 분석을 수행한다. 즉 수집된 HTML의 링크정보를 이용하여 HTML문서를 분석한다. HTML문서를 분석하기 위하여 본 발명에서는 HTML문법에 사용되는 특정한 태그(tag)를 기준으로 문서를 분석하는데 특징이 있다. 예를 들어서 TR, TD 태그를 기준으로 HTML문서를 분석하는 경우에, HTML문서에서 일반 텍스트와 TR, TD태그에 의하여 지정되어 있는 내용만을 추출한다.

도 3a 및 도 3b는 HTML문서를 분석하는 일실시예이다. 도 3a와 같은 HTML문서 데이터에서 태그필터 TR, TD를 적용하여 HTML문서를 필터링하면, 이름, 주소, 홈페이지에 대한 데이터들이 추출된다. 도 3b는 도 3a에 의한 데이터 결과를 간략하게 표시한 도면이다.

상기 도면에서 태그는 추출되지 않았으나, 이것은 적용되는 태그필터의 지정방식에 따라서 변경될 수 있는 것으로서, 만일 태그 tr,td,a를 적용하면 까지 추출되도록 구성할 수 있다. 필터태그를 이용하여 HTML문서를 필터링하여 데이터를 추출하는 경우에 있어서, 모든 속성(Attribute) 값은 무시하는 것이 데이터의 처리가 용이한 이점이 있다. 예를 들어서 <td bgcolor=ffffff textcolor=ff00ff>와 같은 경우에 <td>로만 추출된다.

상기와 같이 HTML문서로부터 태그필터를 이용하여 데이터를 분석, 추출한 후에, 상기 제3단계(S3)로서, HTML문서의 패턴 규칙이 정의된다. 패턴 규칙은 상기 제2단계(S2)에서의 데이터를 정렬하기 위한 것으로서 예를 들면 다음과 같은 형식이 된다.

(1)패턴의 태그는 <tag>와 같이 정의한다.

(2)패턴의 자료는 \$자료명\$과 같이 기술한다.

(3)패턴의 분석시에 링크를 표시하는데 사용되는 태그인 태그는 무시되며, \$URL\$과 같은 패턴 자료를 기술하는 경우에 참조된다.

(4)' '와 ' '는 문서에서 반복되는 패턴을 표시한다. 예를 들어, 도 3a에 표시된 실시예에서,

<tr>

<td>\$이름\$</td><td>\$주소\$</td><td>\$URL\$\$홈페이지\$</td>

</tr>의 경우에는 단 한번만 패턴을 적용한다.

그러나,

<tr>

<td>\$이름\$</td><td>\$주소\$</td><td>\$URL\$\$홈페이지\$</td>

</tr>

의 경우에는 문서에 대하여 반복해서 패턴을 적용한다.

(5)\$URL\$.URL()는 의 '링크' 정보를 추출한다.

(6)\$자료형\$는 다음과 같은 데이터 추출함수에 적용할 수 있다.

\$자료형.RTRIM(n)\$: 텍스트를 오른쪽으로부터 nByte만큼 잘라낸다.

\$자료형.LTRIM(n)\$: 텍스트를 왼쪽으로부터 nByte만큼 잘라낸다.

\$자료형.CRTRIM(c)\$: 텍스트를 오른쪽으로 c문자가 나올때까지 자른다.

\$자료형.CLTRIM(c)\$: 텍스트를 왼쪽으로 c문자가 나올때까지 자른다.

(7)\$URL.SUBURL([필터],패턴)\$와, \$URL.SUBPATTERN([필터],패턴)\$은 다음과 같은 의미를 갖는다. 문법: SUBURL([필터],패턴), SUBPATTERN([필터],패턴)으로서, 차이점은 SUBURL([필터],패턴)은 </a href=>의 링크정보까지 추출하는데 비하여, SUBPATTERN([필터],패턴)은 링크정보를 추출하지 않는다. 상기와 같이 패턴규칙정의에 의하여 변환되어서 추출된 데이터형태가 도 4에 표시되어 있다.

상기 제2단계(S2) 및 제3단계(S3)를 통하여 최종적으로 데이터를 추출하는 것이 제4단계(S4)이다. 추출된 최종데이터의 형태는 도 4에서와 같이, <이름>홍길동, <이름>박길동과 같이 논리적으로 이용하기 용이한 데이터가 되는 것이다.

상기 제4단계(S4)에서는 출력된 데이터를 논리적구조가 개선된 XML구조의 데이터로 변환하는 것이 가능하다. 이것이 단계(S4a)로서 제4단계(S4)에 더 추가될 수 있다. 또는 목적에 따라서 상기 XML 문서로의 변환단계(S4a)를 상기 제4단계(S4)에 포함시켜서 본 발명의 방법을 구성하는 것도 가능함은 명백한 것이다. 다시말하면 XML문서형태가 필요하면 단계(S4a)를 직접 제4단계(S4)에 추가시켜서 구성할 수 있는 것이다.

상기와 같이 추출된 데이터를, 제5단계(S5)에서, 저장매체(24)에 저장한다. 상기 저장매체(24)는 데이터 베이스의 정렬형태로 저장하는 것이 바람직하다. 예를 들어서, 추출된 데이터(XML변환전 또는 변환된 데이터)를 데이터베이스의 테이블과 레코드에 매치할 수 있도록 다음 : (1)자료를 입력할 테이블의 지정, (2)추출된 자료와 필드와의 매치정보와 같이 스키마(schema)를 지정한다.

또한 필드를 매치시키는 규칙은 다음과 같이 지정할 수 있다.

필드타입, 필드명=필드패턴자료

(1)필드타입 : S-문자형자료, N-숫자/날자형 자료.

(2)필드명 : 테이블의 실제 필드명을 지정한다.

(3)필터패턴자료 : 필터패턴형식에서의 \$데이터\$형식이다. 만일 필터패턴자료가 \$데이터\$형식이 아닌 경우, 상수 데이터로 간주한다.

예를 들어서, S.name='홍길동'이면, name 필드에 상수 홍길동을 문자열로 입력한다. 문자열 상수인 경우 반드시 작은따옴표(')와 같은 부호를 붙여서 구별하는 것이 바람직하다. 또한 N.number=1000이면 number 필드에 상수 1000을 숫자로 입력한다.

상기와 같이 구성함으로서 본 발명의 방법에 의하여 웹페이지로부터 데이터가 수집되고, 변환된 후에, 저장매체(24)에 저장됨으로서 분석, 추출, 저장과정이 완료된다. 상기 본 발명의 방법에 의하여 추출된 데이터를 저장하는 저장매체(24)는 또한 다양한 형태의 데이터로 가공, 처리하여 이용할 수 있음은 명백하다.

도 5 내지 도 7a 및 도 7b에는 본 발명의 웹페이지로부터 정보를 추출하고 저장하기 위한 방법 및 시스템을 이용하여 구성된 웹서비스 사이트 구축의 일례가 도시되어 있다. 정보수집모듈(21)과, 필터링 & XML모듈(22)과, DBMS입력모듈(23)과, 저장매체(24)를 포함하고 있으며, 상기 저장매체(24)에 저장된 데이터들을 이용하여 특정한 서비스를 하기 위한 수단들이 더 추가되어 있다.

도 5를 참고하면,전송한 정보수집모듈(21)과, 필터링 & XML모듈(22)과, DBMS입력모듈(23)과, 저장매체로서의 상품DB(24) 이외에, 사용자관리수단(27)과, 검색인터페이스(26)와, 사용자DB(25)와, 검색수단(28)과, 결제관리수단(29)과, 광고 및 푸시관리수단(30)과, 푸시, 이메일(e-mail)서비스수단(31)이 더 표시되어 있다.

상기 상품DB는 본질적으로 본 발명의 웹페이지로부터 정보를 추출하고 저장하기 위한 방법에 의하여 추출된 데이터들이 저장되어 있는 저장매체의 일례로서 동일할 것이다. 상기 상품DB(24)에는 사용자가 특정한 주제에 대하여 수집하고, 추출된 정보가 데이터베이스화되어 저장되어 있으므로 검색수단(28)에 의하여 검색되어서 추출된다.

상기 사용자관리수단(27)은, 사용자의 등록과 로그인을 관리하는 모듈로서, 사용자등록시에 사용자의 등록정보를 입력하게 함으로서 사용자의 성향과 연령, 직업등을 파악하기 위한 기본정보로 활용한다.

상기 검색인터페이스(26)는 사용자가 사이트에 로그인하여 검색수단과 대화하기 위한 통로로서, 다양한 형태의 인터페이스를 생성하여 제공함으로써 효과적인 검색을 지원하도록 하는 역할을 한다.

상기 사용자DB(25)는 사용자관리수단(27)으로부터 전달된 데이터를 보관하는 DB이고, 상품DB(24)는 웹페이지로부터 수집된 데이터중에서 필터링을 통하여 추출된 데이터를 저장하는 DB이다.

상기 검색수단(28)(Information Retrieval System)은 사용자DB와 상품DB의 내용을 필드별로 색인화하여 구조적인 검색이 가능하도록 지원하는 검색수단(시스템)으로서, 각 기능모듈에서 검색수단에 적절한 질의어를 통하여 결과를 요청하고, 요청된 질의어에 대한 결과를 해당 기능모듈(수단)으로 전송하는 역할을 담당하고 있다.

또한 상기 경매관리수단(29)은, DB에 저장된 사용자정보와 상품정보를 활용하여 해당 사용자에게 경매정보를 이메일과 푸시 서비스를 이용하여 제공하고 실제 경매 진행상황을 조정하는 역할을 담당한다.

또한 상기 광고 및 푸시관리수단(30)은, 사용자정보를 활용하여 꼭 필요한 광고만을 사용자에게 제공함으로써 불필요한 광고의 남용을 억제하며, 경매수단(29)과의 연동을 통하여 경매수단에서 공지하고자 하는 내용을 사용자에게 직접 전달하도록 푸시, 이메일 서비스수단(31)에 요청하는 중재역할을 담당한다.

또한 상기 푸시, 이메일서비스수단(31)은 광고 및 푸시관리수단의 요청에 의하여 해당 사용자에게 직접 제공하거나 이메일을 통하여 제공하는 역할을 한다.

도 6은 여행정보를 검색하기 위한 사용자인터페이스의 일례를 도시한 것으로서, 여행목적지를 선택하기 위하여, 대륙, 국가, 도시등을 선택하기 위한 메뉴가 구비되어 있으며, 검색조건을 입력하기 위한 선택메뉴가 표시될 수 있다. 이러한 메뉴화면은 다양한 형태로 구성이 가능함은 명백하다.

도 7a 및 도 7b는 도 6을 통하여 입력한 정보에 따라서 웹페이지들을 검색하고 최종적으로 추출한 데이터가 표시되어 있다.

도 7a는 다수의 여행정보제공 웹페이지가 표시되어 있으며, 이러한 웹페이지로부터 본 발명의 방법에 의하여 여행정보에 대한 데이터가 추출되어서 상품DB(24)에 저장된다.

도 7b는 상품DB(24)로부터 추출된 여행정보데이터를 일목요연하게 화면상에 표시한 것으로서, 불필요한 정보대신에 사용자가 꼭 필요로 하는 필수정보만이 표시되므로 사용자는 신속하게 여행지와 여행조건들을 확인하고 선택할 수 있다. 따라서 종래 검색방식에 비하여 시간과 속도가 단축되는 등의 효과가 있다.

즉, 최종추출된 자료는 하나의 여행상품에 대한 것이 아니라, 다양한 상품들이 서로 비교될 수 있기 때문에 사용자는 하나의 화면 또는 소수의 화면에서 각 상품들을 비교하면서 적절한 상품을 선택할 수 있는 특징이 있다.

상기 설명에서는 여행상품에 대한 정보추출방식을 예로 들었지만, 가전제품, 전자제품, 전기용품등과 같은 다양한 상품들도 역시 동일하게 적용될 수 있음은 명백한 것이다.

발명의 효과

상기와 같이 본 발명에 의하면, HTML문서의 비논리적인 구조중에서 특정한 패턴(pattern)을 가진 부분을 처리함에 있어서 범용성을 가지는 웹페이지로부터 정보를 추출하고 저장할 수 있는 이점이 있다.

또한 패턴분석을 통하여 추출되는 정보를 논리적인 구조로 추출해 냄으로서 보다 유용하게 정보를 처리할 수 있는 이점이 있다.

또한 데이터베이스와의 연동을 통하여 정보의 재활용을 용이하게 할 수 있는 이점이 있다.

더욱이, 웹페이지로부터 정보를 추출하고 저장하기 위한 본 발명의 방법을 이용하여 다양한 형태의 사업성있는 서비스모형을 제공함으로써 정보획득의 신속, 정확성을 제고하여 경쟁력을 높일 수 있는 이점이 있다.

본 발명은 기재된 구체예에 대해서만 상세히 설명되었지만 본 발명의 사상과 범위내에서 변형이나 변경할 수 있음은 본 발명이 속하는 분야의 당업자에게는 명백한 것이며, 그러한 변형이나 변경은 첨부한 특허청구범위에 속한다 할 것이다.

(57) 청구의 범위

청구항 1. 다수의 클라이언트PC와, 다수의 웹서비스제공용 서버와, 적어도 하나 이상의 웹페이지정보를 추출하기 위한 서버(6)를 포함하는 인터넷 통신시스템에서,

웹상의 HTML 문서의 링크정보를 수집하여 문서의 정보를 구축하는 제1단계와, 수집된 HTML문서의 링크정보를 이용하여 특정태그를 기준으로 HTML문서를 분석하는 제2단계와, 분석된 HTML문서에 대한 HTML문서의 패턴을 정의하는 제3단계와, 상기 제3단계에서 정의된 패턴에 따라서 데이터를 추출하는 제4단계와, 추출된 데이터를 데이터베이스형태로 저장하는 제5단계를 포함하는 것을 특징으로 하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법.

청구항 2. 제1항에 있어서,

상기 제4단계에서 패턴에 따라서 추출된 데이터를 XML문서로 변환하는 단계(S4a)를 더 포함하는 것을 특징으로 하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법.

청구항 3. 제1항에 있어서,

상기 제1단계에서 문서의 구축이 사용자가 자신이 찾기 원하는 정보를 포함하는 기준 URL을 입력하면 이것을 기점으로 하여 웹페이지에 링크된 웹사이트를 따라가며 HTML문서의 링크정보를 수집하는 것을 특징으로 하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법.

청구항 4. 제1항에 있어서,

상기 제2단계에서 태그필터를 이용하여 HTML문서를 필터링하는 경우에 속성값이 무시되고 추출되지 않는 것을 특징으로 하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법.

청구항 5. 제1항에 있어서,

상기 제3단계에서의 패턴정의규칙이,

(1)패턴의 태그는 <tag>와 같이 정의하며,

(2)패턴의 자료는 \$자료명\$과 같이 기술하며,

(3)패턴의 분석시에 링크를 표시하는데 사용되는 태그인 태그는 무시되고, \$URL\$과 같은 패턴 자료를 기술하는 경우에 참조되며,

(4)' '와 ' '는 문서에서 반복되는 패턴을 표시하며,

(5)\$URL\$.URL()\$는 의 '링크' 정보를 추출하며,

(6)\$자료형\$로 표시되는 형들로서, \$자료형.RTRIM(n)\$는 텍스트를 오른쪽으로부터 nbyte만큼 잘라내고, \$자료형.LTRIM(n)\$는 텍스트를 왼쪽으로부터 nbyte만큼 잘라내며, \$자료형.CRTRIM(c)\$는 텍스트를 오른쪽으로 c문자가 나올때까지 자르며, \$자료형.CLTRIM(c)\$는 텍스트를 왼쪽으로 c문자가 나올때까지 자르는 것과 같이 데이터 추출함수에 적용할 수 있으며,

(7)\$URL.SUBURL([필터],패턴)\$와, \$URL.SUBPATTERN([필터],패턴)\$은, 문법: SUBURL([필터],패턴), SUBPATTERN([필터],패턴)의 형태로 사용되며, SUBURL([필터],패턴)은 의 링크정보까지 추출하고, SUBPATTERN([필터],패턴)은 링크정보를 추출하지 않도록 정의되는 것을 특징으로 하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법.

청구항 6. 제1항에 있어서,

상기 제5단계에서의 데이터베이스 저장방식이,

추출된 데이터(XML변환전 또는 변환된 데이터)를 데이터베이스의 테이블과 레코드에 매치할 수 있도록 (1)자료를 입력할 테이블의 지정, (2)추출된 자료와 필드와의 매치정보'를 포함하도록 스키마를 지정하며,

또한 필드를 매치시키는 규칙이,

'필드타입,필드명=필드패턴자료'로서

(1)필드타입 : S-문자형자료, N-숫자/날자형 자료.

(2)필드명 : 테이블의 실제 필드명, 그리고

(3)필드패턴자료 : 필터패턴형식에서의 \$데이터\$형으로서, 필터패턴자료가 \$데이터\$형식이 아닌 경우, 상수 데이터로 간주하도록 지정하는 것을 특징으로 하는 웹페이지로부터 정보를 추출하고 저장하기 위한 방법.

청구항 7. 다수의 클라이언트PC와, 다수의 웹서비스제공용 서버와, 적어도 하나 이상의 웹페이지정보를 추출하기 위한 서버(6)를 포함하는 인터넷 통신시스템에 있어서,

상기 서버(6)가, 인터넷에 산재한 웹페이지중의 특정주제를 포함한 웹페이지로부터 정보를 수집하는 정보수집모듈(21)과, 상기 정보수집모듈(21)에 의하여 수집된 정보로부터 유용한 형태로 정보를 추출하기 위한 필터링 & XML모듈(22)과, 상기 필터링 & XML모듈(22)로부터 출력되는 데이터를 저장매체(24)에 저장하기 위한 DBMS입력모듈(23)을 포함하는 것을 특징으로 하는 웹페이지로부터 정보를 추출하고 저장하기 위한 시스템.

청구항 8. 제1항 내지 제6항의 방법에 의하여 처리된 웹페이지의 정보를 저장하는 저장매체.

청구항 9. 제1항 또는 제8항의 저장매체를 포함하며, 상기 저장매체에 저장된 데이터를 처리하여 이용하도록 구성되는 웹사이트 서비스 시스템.

청구항 10. 다수의 클라이언트PC와, 다수의 웹서비스제공용 서버와, 적어도 하나 이상의 웹페이지정보를 추출하기 위한 서버(6)를 포함하는 인터넷 통신시스템에 있어서,

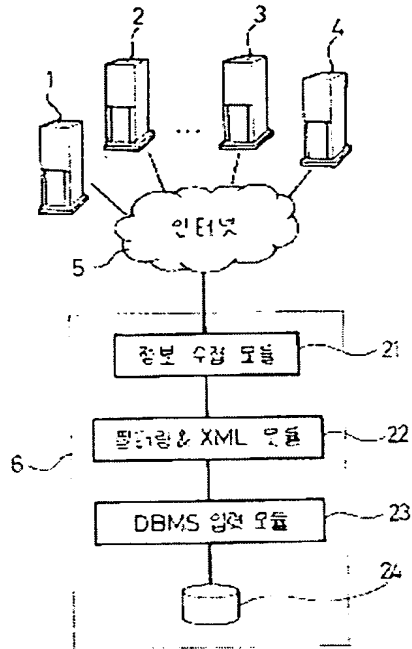
상기 서버(6)가,

인터넷에 산재한 웹페이지중의 특정주제를 포함한 웹페이지로부터 정보를 수집하는 정보수집모듈(21)과, 상기 정보수집모듈(21)에 의하여 수집된 정보로부터 유용한 형태로 정보를 추출하기 위한 필터링 & XML모듈(22)과, 상기 필터링 & XML모듈(22)로부터 출력되는 데이터를 저장매체(24)에 저장하기 위한 DBMS입력모

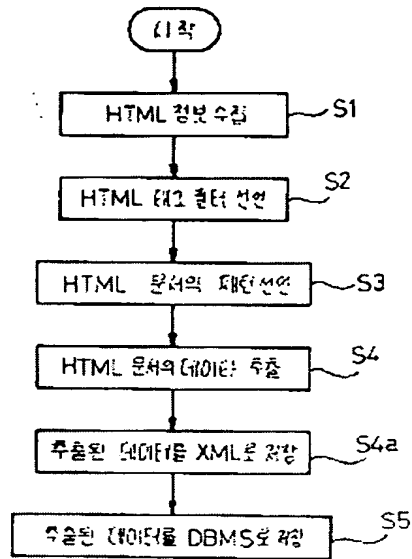
들(23)과, 사용자의 등록과 로그인을 관리하기 위한 사용자관리수단(27)과, 사용자가 사이트에 로그인할 때 대화동로역할을 하는 검색인터페이스(26)와, 각 DB에 저장된 데이터를 필드별로 색인하여 검색이 가능하도록 지원하는 검색수단(28)과, 상품정보와 사용자정보를 활용하여 경매정보를 제공하는 경매관리수단(29)과, 사용자정보를 활용하여 필요한 광고를 제공하는 광고 및 푸시관리수단(30)과, 광고 및 푸시관리수단(30)의 요청에 의하여 해당 사용자에게 정보를 직접 제공하는 푸시, 이메일서비스수단(31)을 포함하는 것을 특징으로 하는 웹서비스 시스템.

도면

도면1



도면2



도면3a

```

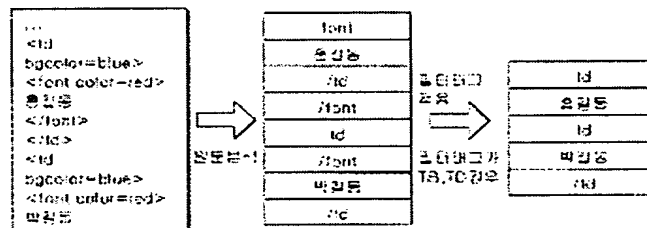
<html>
<body>
<table>
<tr><td>이름</td><td>주소</td><td>휴대폰</td></tr>
<tr><td bgcolor=#0000FF>이름1</td><td>주소1</td><td><a href=home1.html>홈페이지1</a></td></tr>
<tr><td></td>
<tr><td bgcolor=#0000FF>이름2</td><td>주소2</td><td><a href=home1.html>홈페이지2</a></td></tr>
<tr><td></td>
<tr><td>이름3</td><td>주소3</td><td><a href=home1.html>홈페이지3</a></td></tr>
</table>
</body>
</html>
    
```

태그 필터 TR, TD 내용 필터링

```

<tr><td>이름</td><td>주소</td><td>휴대폰</td></tr>
<tr><td>이름1</td><td>주소1</td><td><a href=home1.html>홈페이지1</a></td></tr>
<tr><td>이름2</td><td>주소2</td><td><a href=home1.html>홈페이지2</a></td></tr>
<tr><td>이름3</td><td>주소3</td><td><a href=home1.html>홈페이지3</a></td></tr>
    
```

도면3b



6253 북한신문 기사

547

圖 2-10 鋼筋的斷面

1. 2. 3.

100

 100
 100

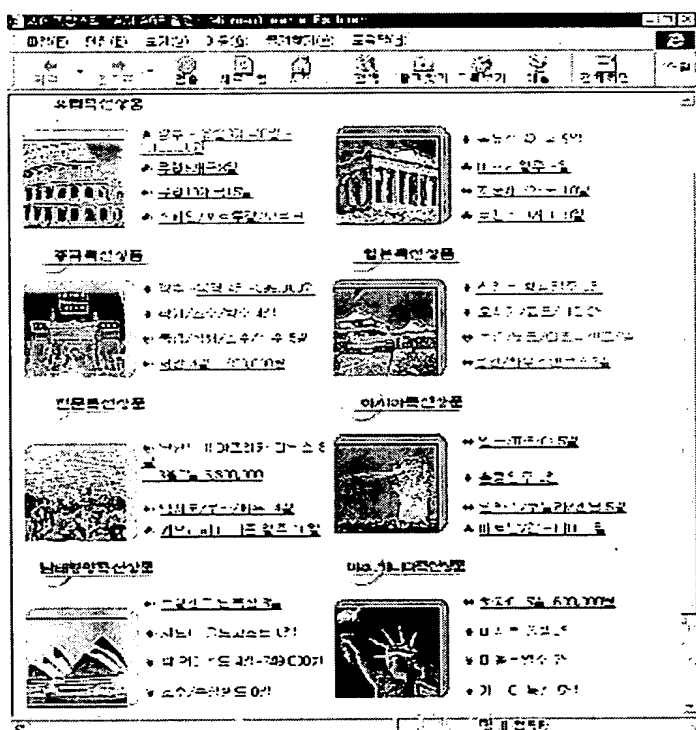
• 1970 •

[illegible]

www.dhammadownload.com

정석포럼 4가,

CP78



B	delimited.XLS	H22	B22
000001	000001	000001	000001
[xml version="1.0" encoding="EUC-KR" standalone="no"]			
[DOCUMENT]			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			
[<수출>] 수출 3개국 - 1,180,000			